Develop Machine Leaning Algorithms to Predict Online Bids Are Made by Robot or Human

CSI5155
YINGJIN GUAN
7385769
GROUP 35

Abstract. -Due to the popularity of using online platform with auction and development of AI technology, there are some concern about AI technology is used to win the online bid over human control. Machine learning can used exist data to analyze if the online bid is made by human or robot so that this situation is prevented in the future in order to keep fairness to human users. This project proposes a study about implementation of machine learning algorithms with online dataset to determine if the auction is placed by robot or human and analyze bidder information and bid information to observe online auctions and learn bidding behaviors in order to classifier if the online bid is made by robot or human.

I. INTRODUCTION

Online auction is known as a popular service for selling and bidding of products or service by users in the internet nowadays. Buyer and seller in different locations in the world can use the platform provided by online auction website or applications to trade. Bidding is a different from fixed price product trading of online platform. Fixed price products can be bought by easily clicking “buy” online and the purchase price will be the price listed beside the product, however for auction, that is not how it works. For example, in Ebay, a famous online shopping site known by auctions, buyers need to place competing bids that higher than the current bid, and after a certain time slot, the auction will end, the winner with the highest bid will get this product.

Artificial intelligence was firstly studied for maximizing the probability of success by using behavior and environment [1]. Following by the AI and robotics development, there are more and more possibilities for robot involve human’s lives. Since AI is playing a significant role involving machine learning algorithm, AI development can use human behavior and environment elements to improve the logical and mathematical knowledge of computer. For example, it can used in games to develop non player character to assistant true human player in the computer game or used in household as smart home devices [2] it is It is oblivious fact that robot technology can be used in various ways to compete with human. Robotics has found to in most studies that used in map exploration [3] and Multi-agent coordination techniques. [4] Using AI technology to place bid in online auction and it can increase probability to win because comparing to real human controlling, robot has much faster reaction to change the bid information when condition changing, this would be not fair to other competitors. For example, robot can even place a bid 0.01 second before the bid closed, and other users will lose chance to win the bid. Also, in the newspaper [5], issues are noticed that robot bidder technology used on penny auction sites to win the free auction, which brings unfair treatment against human users who depend on the manual bid. It is important to study relationship between online bid and bidder since more and more people are using online platform to trade.

Observing the dataset, we can see that it is supervised-learning project since the classes of outcome is labeled as 1 or 0 to represent robot or human. It is binary class study with 10 features

from bid and bidder information. Regression and Classification algorithms are used usually in supervised learning project, linear regression is basic to use find relationship among the outcome class and given features. Thus, for this project the linear classifier will be the first to use to see if the relationship between outcome and features exist linear relationship. However, when observing the data, it is obvious that the evaluation can be predicted to be low since the dataset involve multiple complicated features; To study with more precise result, this project will focus on using different algorithms to study, which are Naïve Bayes, K-Nearest Neighbor, a rule-based algorithm, decision tree algorithm and random forest algorithm.

The project aims to determine whether there are any strong relationships can be defined from the bid information so that it can used to predict if the bid is made by robot or human. Throughout the project, different machine learning algorithms will be used for the dataset, comparisons with precision, recall, accuracy, ROC curve, and AUC values, running time of algorithms will be learned and discussed with results for decision making.

II. CASE STUDY

Robotics are studied with artificial intelligence in e-commerce, as for bidder system, robotics will be useful to increase the probability to win the bid. On the other hand, studying AI bidders is important for online auction platform to balance the fairness between human user and robotic bidders. As a result, studying bidder behavior and environment element is crucial for online auction system.

Online auctions were studied with duration distributions and bid arrivals patterns in websites in order to find bidding strategies. [6] Sampled eBay auctions are resulted for 7-day eBay auction is most popular, duration preference is different from different regions. Also, the last-minute bidding has high proportion bidder in UK. This study can be useful for the future bidding strategy for robotic bidding detection system.

Bidding behavior in the pay-per-bid auction and profits in auctions are studied with 2-step cluster algorithm to analyze irrational bidders and rational bidders.[7] The paper has studied the bidding behavior and calculated the winning probabilities and the expected profits of the bidders to help future bidders to consider how to bid.

Smart bidding strategy of demand side loads involving reinforcement learning is also studied.[8] the project built a piecewise regulation cost model for the trading system. At the same time, they used trading result and reinforcement learning study to analyze bidding acceptance probability. The study has improved decision behavior of loads. This paper has provided a method to improve bidding strategy and benefit online bidding market demand elasticity.

Bidding robot is developed with multi-agent cooperative bidding mechanism. [9] The product introduced can help human bidder to attend, monitor and bid in multiple online auctions. It can even collect information about the auction product with market price. In addition, there is a paper studied to coordinate mobile robot to using “graph-cut” strategies to handle multiple auctions.[10]

Machine learning methods are widely used to develop artificial intelligence, following by development of online auctions, and it is predictable that more and more robotic system like bidding bot will be developed as competitive products against human users. To study the behavior of bidders can be helpful for online auction platform to detect robot bidders. The dataset of the project is consisting of several features of bidder and also features of bid placed by different bidders. Moreover, the robot bidders are labeled, supervised learning algorithms will be used to find how these features can help to detect robot bidder when analyzing characteristic of bid.

III. EXPERIMENTAL SETUP

MySQL is used to observe dataset information, since the dataset has 7.6 million entries, MySQL is also used to reduce dataset size. Jupyter Notebook is used for the rest part of this project. Python is used in the Jupyter Notebook for the programming language for the data preprocessing and algorithm implementation. The git repository is created in the following URL: https://github.com/ginguan/MLProject.git.

IV. DATA OVERVIEW

Original dataset has 7.6 million entries, and bidder information with labeled outcome are shown in train.csv file and information of bid placed by bidders are shown in bids.csv file.

Table 1. Bidder information (train.csv)

Bidder_id	Id of bidder
Payment_account	Payment account of the bidder
Address	Address of bidder
Outcome	Robot or human label for the bidder

Figure 1. Bidder information sample

bidder_id	payment_account	address	outcome
91a3c57b13234af24875c56fb7e2b2f4rb56a	a3d2de7675556553a5f08e4c88d2c228754av	a3d2de7675556553a5f08e4c88d2c228vt0u4	0.0
624f258b49e77713fc34034560f93fb3hu3jo	a3d2de7675556553a5f08e4c88d2c228v1sga	ae87054e5a97a8f840a3991d12611fdcrfbq3	0.0
1c5f4fc669099bfbfac515cd26997bd12ruaj	a3d2de7675556553a5f08e4c88d2c2280cybl	92520288b50f03907041887884ba49c0cl0pd	0.0
4bee9aba2abda51bf43d639013d6efe12iycd	51d80e233f7b6a7dfdee484a3c120f3b2ita8	4cb9717c8ad7e88a9a284989dd79b98dbevyi	0.0
4ab12bc61c82ddd9c2d65e60555808acqgos1	a3d2de7675556553a5f08e4c88d2c22857ddh	2a96c3ce94b3be921e0296097b88b56a7x1ji	0.0

Table 2. Bid information (bids.csv)

Bid_id	Id of bid
Bidder_id	Id of bidder
Auction	Id of auction
Merchandise	Category of the auction site campaign, search term or online advertisement, for example, mobile, jewelry
Device	Phone model of visitor
time	Time of the bide made
Country	Country that IP belongs to
IP	IP address of bidder
URL	URL where bidder from

Figure 2. Bid information sample

bid_id	bidder_id	auction	merchandise	device	time	country	ip	url
0	8dac2b259fd1c6d1120e519fb1ac14fbqvax8	ewmzr	jewelry	phone0	9759243157894736	us	69.166.231.58	vasstdc27m7nks3
1	668d393e858e8126275433046bbd35c6tywop	aeqok	furniture	phone1	9759243157894736	in	50.201.125.84	jmqlhflrzwuay9c
2	aa5f360084278b35d746fa6af3a7a1a5ra3xe	wa00e	home goods	phone2	9759243157894736	py	112.54.208.157	vasstdc27m7nks3
3	3939ac3ef7d472a59a9c5f893dd3e39fh9ofi	jefix	jewelry	phone4	9759243157894736	in	18.99.175.133	vasstdc27m7nks3
4	8393c48eaf4b8fa96886edc7cf27b372dsibi	jefix	jewelry	phone5	9759243157894736	in	145.138.5.37	vasstdc27m7nks3

V. DATA PREPROCESSING

After observing the original data, for feature engineering to analyze raw dataset, missing value is detected and found in 15 columns in “country” in bid dataset, which is only 0.00000195916 of the datasets, therefore, this problem is ignore. When combing the bid dataset and bidder dataset, missing value is 29 for column ‘country’ and 36 for other columns, the ratio is also very low, thus, these rows with missing values are dropped. join the 2 tables together using SQL script.

Additionally, it is important to notice that there are only 47559 rows of the class 1 represent robot, and others are class 0 represent human so that the data set is extremely imbalance. Then, the performance of machine learning classifier will be affected. Majority class will show large

difference in confusion matrix and also other evaluation attributes. Under this situation, an easy way to balance the data is choosing similar amount of class 0 and join to the class 1 data (detail code is located in sqlscript.sql file). Fist select all rows for outcome class “1” from the dataset. Then, choose certain amount of outcome class “0” to join the class “1” using join method in SQL. After processing with MySQL, 66601 of class 0 is chosen to join class ‘1’, as “select.csv” in file. (total 114148 rows).

Table 3. Same bidder_id sample

bidder_id	payment_account	address
e12177ff9c1a8413996f7b1a590980c82ofeo	4ceb0ecc3a5bf6988c1639c112ea88eaoit7w	acb0169abfffe7093ba53f8a4472ff4b14lk7
e12177ff9c1a8413996f7b1a590980c82ofeo	4ceb0ecc3a5bf6988c1639c112ea88eaoit7w	acb0169abfffe7093ba53f8a4472ff4b14lk7
e12177ff9c1a8413996f7b1a590980c82ofeo	4ceb0ecc3a5bf6988c1639c112ea88eaoit7w	acb0169abfffe7093ba53f8a4472ff4b14lk7
e12177ff9c1a8413996f7b1a590980c82ofeo	4ceb0ecc3a5bf6988c1639c112ea88eaoit7w	acb0169abfffe7093ba53f8a4472ff4b14lk7

As we can see from Table 3, it is discovered that when bidder_id is the same, the ‘payment_account’ and ‘address’ values remain the same. This represents all these three features are having same effects or relationship for the outcome, therefore, only bidder_id column is kept, and the other 2 columns can be dropped.

For the ‘merchandise’ feature, there are 10 categories: 'furniture', 'mobile', 'sporting goods', 'home goods', 'jewelry', 'office equipment', 'computers', 'books and music', 'auto parts', 'clothing'. Use rank and normalize based on common sense assumption for price level, the ranking is from 1 to 10 and normalize to 0.1 to 1. (code located in “select.ipynb”) as a result, for example jewelry which is considered as the most expensive category will be rank as 10 and after normalization, 1 is the value representing the jewelry category in all dataset.

Table 4. Rank and Normalization of Merchandise column.

books and music	office equipment	computers	clothing	Home goods	furniture	Sporting goods	Auto parts	mobile	jewelry
1	2	3	4	5	6	7	8	9	10
0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1

For the ‘country’, there are 198 countries in with 2 character labeling. **labelEncoder()** is used to transform the 198 countries in range from 0 to 197. Result is represented in Figure 3 and 4. For example, the country labeled as “us” is transformed to 84 as integer value.

Figure 3. Overview of country values

```
array(['us', 'in', 'py', 'ru', 'th', 'id', 'za', 'ng', 'sd', 'au', 'hr',
       'np', 'iq', 'bd', 'tr', 'ch', 'ke', 'uk', 'fr', 'pk', 'my', 'vn',
       'ro', 'gh', 'ua', 'pl', 'by', 'ar', 'zm', 'lk', 'ph', 'br', 'es',
       'mx', 'il', 'qa', 'nl', 've', 'sg', 'gt', 'ae', 'az', 'uz', 'ht',
       'tz', 'gm', 'dk', 'no', 'kw', 'mk', 'hu', 'it', 'ml', 'sv', 'bn',
       'ni', 'cn', 'et', 'ge', 'mw', 'ee', 'ye', 'kr', 'tn', 'gr', 'at',
       'cm', 'ca', 'mn', 'rs', 'sz', 'pe', 'jp', 'sl', 'bh', 'zw', 'bg',
       'de', 'eu', 'cr', 'jo', 'ie', 'sa', 'eg', 'dz', 'hk', 'ec', 'si',
       'lv', 'na', 'mt', 'ug', 'kg', 'se', 'bb', 'sc', 'sn', 'om', 'fi',
       'cl', 'ma', 'am', 'lr', 'be', 'bf', 'kh', 'md', 'ly', 'al', 'ba',
       'bo', 'lt', 'ga', 'mr', 'jm', 'bj', 'mu', 'pa', 'cz', 'ao', 'lu',
       'me', 'af', 'kz', 'hn', 'ls', 'uy', 'lb', 'cy', 'sk', 'ir', 'la',
       'dj', 'bz', 'ci', 'is', 'mg', 'so', 'co', 'pt', 'gy', 'td', 'rw',
       'pr', 'bw', 'gq', 'cv', 'mc', 'ne', 'tg', 'bi', 'sy', 'tt', 'cd',
       'sb', 'mz', 'mm', 'tj', 'tw', 'gu', 'cg', 'gl', 'nz', 'mv', 'ps',
       'tm', 'ag', 'ad', 'sr', 'ws', 'je', 'do', 'li', 'fj', 'nc', 'gi',
       'cf', 'mo', nan, 'dm', 'bt', 're', 'fo', 'mp', 'bm', 'gn', 'tl',
       'pg', 'pf', 'vc', 'zz', 'bs', 'aw', 'gb', 'vi', 'mh', 'tc', 'an',
       'er', 'gp'], dtype=object)
```

Figure 4. Transformation of Country value

```
array([ 84,  15,  81,  25, 132, 191, 144, 154, 174, 186,  31,  41,  65,
        93, 156, 182, 102,  23, 143,  75, 127, 108, 194, 179,  83,  95,
        19, 159, 133, 100,  86,   8, 126,  53, 161,   1, 178, 196, 185,
        76,  90, 152,  56,  39, 109,  55, 153,  85,  91, 145, 189, 188,
       187, 128,  88, 150,  12, 135,  37, 103, 125,  57, 138,  99, 124,
        44, 123, 193, 148,  78, 169,  97, 139, 195, 184,   9,  79, 155,
        59, 110,  38,  36, 183,  45,  32, 149,  13,  72,  24,  46,  77,
        16, 136,  73,  14, 168,  94,  58,  98,  29,  40, 116,  10, 177,
       170, 181, 157,  64,  21,  50,  61, 164, 104, 160, 163, 111, 165,
        17, 158,  42,  51,  80, 113, 134, 118,   5,   4, 121, 162,  27,
       105, 129,  34,  43,  68, 192, 180, 131,  82,  62, 166,  96,   2,
       107, 175, 140,   7,  47, 115,  52, 173, 112, 142,  18,  35, 117,
       106, 122,   0,  28,  71,  87, 137, 172,  92,  20,  49, 146,  74,
        67,  33, 167, 114,  63, 120, 141, 147,  30,  22, 176, 101,  69,
       119, 197,  89,  48,  26,  11,   3, 190, 130, 151, 171,  60,  70,
         6,  66,  54])
```

Then, it is observed that 'time' value should be normalize since range is [9631916842105264, 9772885210526316] and result to [0,100]. Column '**device'** is found that all values are 'phone'+ numeric string value, therefore, the values in device are subtracted string "phone" and converted to leave only the numeric part.

labelEncoder() is also used to transform "bid_id", "bidder_id", "auction", "ip", "url".

Also after the other non- numerical data except bid_id and bidder_id are transferred to numerical data, use **preprocessing.MinMaxScaler(feature_range=(0,100)),** this method follows the formula:

$$\text{NewValuei} = \frac{\text{xi} - \min(\text{x})}{\max(\text{x}) - \min(\text{x})} * 100$$

to normalize data into range [0,100]. Table 5 shows the data sample after normalization and transformation. Now the dataset are processed and prepared for the model construction.

Table 5. Data Processed Overview.

	bid_id	bidder_id	auction	merchandise	device	time	country	ip	url	outcome
0	20	365	55.188367	0.2	0.475238	0.002292	57	19.645109	86.979313	1
1	36	183	3.040317	0.9	0.825413	0.003438	147	61.384112	30.403445	1
2	47	158	14.111038	0.9	1.100550	0.005730	62	97.832642	12.200069	1
3	55	368	34.434898	0.5	0.100050	0.006877	151	47.497305	86.979313	1
4	107	447	43.258427	0.7	2.026013	0.012607	146	96.730688	11.479297	1

V. MODEL CONSTRUCTION AND EVALUATION

Model construction uses scikit-learn library with python. 5 types of models are used: linear classifier, tree-based model, distance-based, rule-based and ensemble model. For each model, use 30% for testing dataset, and 60% for training dataset when dividing training dataset and testing set in code, the training set and testing set will be chosen randomly. Accuracy, recall, precision, confusion matrix and ROC curve are used for evaluation since they are good attributes for the evaluations of algorithms for data. Accuracy is ratio of correctly predicted data to the whole data. Recall is the ratio of correctly predicted data to the actual positive class. Precision is the ratio of correctly predicted positive class to the whole predicted positive class. ROC (reciver operating characteristics) curve is used to illustrate the relationship between true positive rate and false positive rate. When the curve is closer to the TPR axis and also the top border of the graph, the test is more accurate. AUC is the area under ROC curve, which shows the larger value will come with more accurate testing. At the end, the comparison of evaluation attributes will be used to compare the general performance of algorithms in this project.

5.1 Linear classifier

Linear classifier is used as SGDClassifier() which is stochastic gradient descent, from sklearn.linear_model. SGD is discriminative learning method of linear classifiers under loss function, and loss function is chosen by "hinge", which is Support Vector Machines. SGD is measured with 2 arrays, an array of size with training samples, and array of size with outcome values. And the coefficient for each class is shown as following for the linear model. Coefficient for each class is calculated to show how to draw the linear for the linear model.

```
Coefficient for each features:
 [[  -2495.49329112 -683216.7664553  -102776.08953261   -1572.55926135
    -8493.14771089   18078.75426257 -181322.2009718  -109611.39584306
  -152571.75263981]]
```

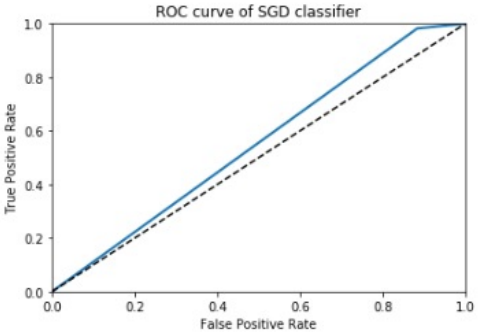
Figure 8. coefficient for each class

Results are different each time run but the coefficient for each feature are similar, 2 samples result of coefficient for the linear model are shown in Figure 8.

Figure 9. Confusion matrix for linear

```
==================
Confusionn matrix
[[ 2336 17735]
 [  252 13922]]
==================
```

Figure 10. ROC curve for linear



AUC: 0.549303897367253

For evaluation, accuracy=0.47, recall=0.47, precision=0.71, confusion matrix, and AUC value are calculated. Confusion matrix and ROC curve with AUC value are shown in Figure 9 and Figure 10. Linear regression algorithm uses stochastic gradient descent, SGD is well- applied to large size dataset and more than 10^5 features. However, the algorithm is limited when non-linear relationship occurs. Although running time is 0.1304, which is short., but other evaluation attributes show the classifier is not ideal.

Dataset of bid are complex relationship; thus, the performance is poor. Accuracy recall show evaluation are under 50%. From this basic model, it illustrates that the relationship of this project is non-linear and more work need to improve the model.

5.2 Naïve Bayes

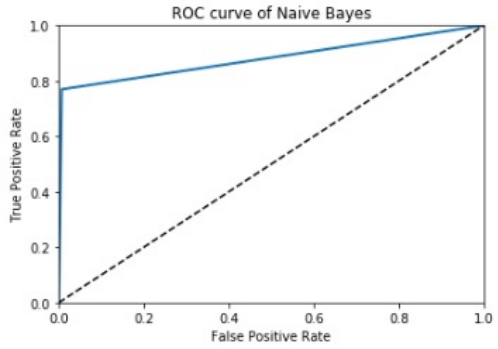
Naïve Bayes is used as next because it can perform better than linear model. Therefore, it is chosen., algorithm calculate the posterior probability for each class and choose the higher posterior probability to be result the prediction. Naïve Bayes algorithm is more efficient than linear regression, it can perform well with outlier, however it is not good for large dataset, so the estimated probabilities are not really reliable than the predicted values. For the project, Gaussian Naïve Bayes algorithm is chosen and the likehood function (Eq.1) is using as illustrated by the scikit learn library.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$ Eq.1

Figure 12. *Evaluation Attribute Naïve Bayes*

```
==================
Confusionn matrix
[[19672   138]
 [ 3322 11113]]
==================
```

Figure 11. ROC curve for Naïve Bayes




```
AUC:  0.8814493664876947
```

For evaluation, accuracy=0.89, recall=0.89, precision=0.91, confusion matrix, and AUC value are calculated. Confusion matrix and ROC curve with AUC value are shown in Figure 11 and Figure 12. The running time is 0.045, which is fast comparing to linear model. The performance of Naïve Bayes is not really good, but it can be considered to use when time is a primary element for decision making.

5.3 K-Nearest Neighbors Algorithm

K-Nearest Neighbors algorithm is used as distance-based algorithm. Weight function is choosing "distance", which uses closer neighbors of points with larger effects than neighbors with further distance.

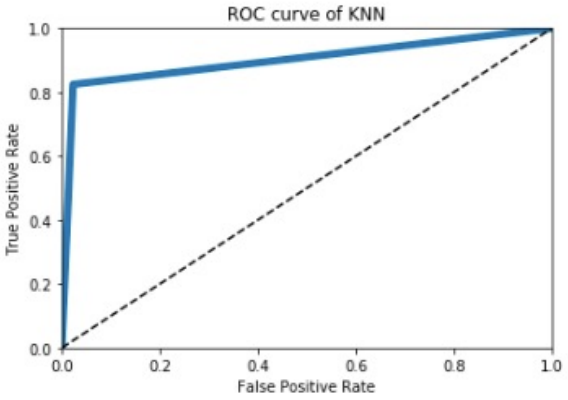
Value of k is the first thing to calculate since different value of k will affect the performance of algorithm. According to the class note, suitable value of k can be chosen when observing the lowest error rate of testing data and keep the k value not too large. To determine the suitable value of k for the dataset, use k = [1,40] to test the error rate for testing dataset to choose the suitable value for k, therefore, k = 15 is chosen and also k=6 is chosen because it locates at the "elbow" part of the error rate of testing data.

Figure 14. *Evaluation Attribute KNN (k=6)*

```
Accuracy of KNN: 0.9128339903635567
Recall of KNN: 0.9128339903635567
Precision of KNN: 0.9176307881986377

==================
Confusionn matrix
[[19357   453]
 [ 2532 11903]]
==================
```

Figure 13. ROC curve for KNN



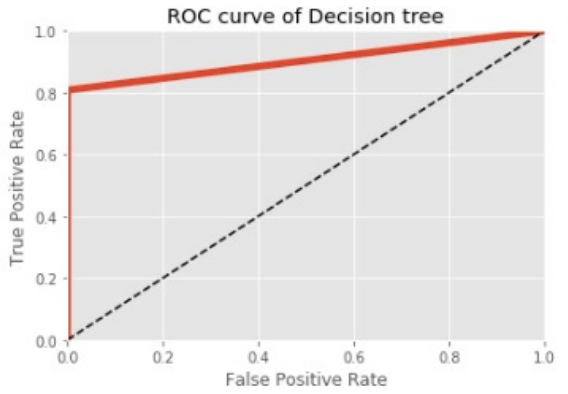

```
AUC:  0.900862882174562
```

Figure 16. *Evaluation Attribute KNN (k=15)*

```
Accuracy of KNN: 0.9192874872244123
Recall of KNN: 0.9192874872244123
Precision of KNN: 0.9283194570748496

==================
Confusionn matrix
[[19758    52]
 [ 2712 11723]]
==================
```

Figure 15. ROC curve for KNN




```
AUC:  0.9041219258746103
```

For both evaluation, accuracy=0.91, recall=0.91, precision=0.92, confusion matrix, and AUC value are calculated. Consider with running time, when k = 6, running time is 0.68, when k = 15, running time is 0.73. From the result above, it is obvious that the evaluation went better with KNN algorithm comparing to the linear model and Naïve Bayes. The result evaluation for k = 6 and k = 15 are similar. In order to get shorter running time, k = 6 will be more suitable for the dataset.

Using KNN without testing value of k is considered when doing the project, however, in this case, most suitable value of k will not be determined, and small value of k may be used to avoid higher variance. As a result, checking the lowest error rate of testing data is not kept to find value of k.

K-nearest neighbor is effective if training dataset is large enough and it can be efficient when the dataset is already preprocessed, therefore, the accuracy is over 90%. However, in order to choose value of k, computation cost is high, and also distance of instance and the calculation time is longer than other algorithms.

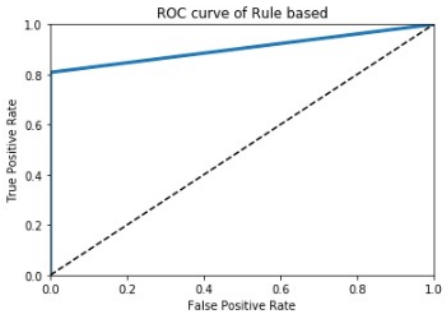
Rule based algorithm

Skope- rule is used for constructing model, this module built for learning logical rules for "scoping" the target class, "outcome" in dataset. In this case, the setting is by default of minimum precision = 0.3, minimum recall = 0.1.

Figure 17. *Evaluation Attribute Rule-based*

```
==================
Confusionn matrix
[[19810     0]
 [ 2766 11669]]
==================
```

Figure 18. ROC curve for rule-based



AUC: 0.9041912019397298

For evaluation, accuracy=0.92, recall=0.92, precision=0.93, confusion matrix, and AUC value are calculated. Confusion matrix and ROC curve with AUC value are shown in Figure 17 and Figure 18. The classifier shows nice performance to predict positive class of the data. However, the running time of this algorithm is 42.3, which is too long and computation cost is too large. And the performance of overall classifier is slightly better than K-nearest neighbor algorithm.

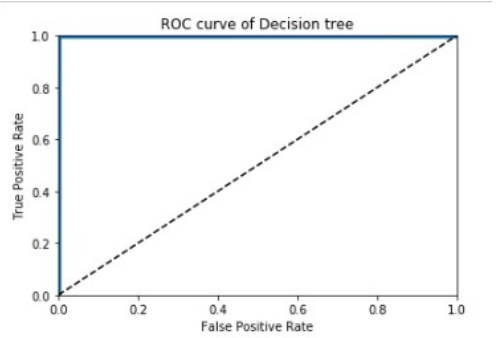
Decision tree algorithm

Decision tree algorithm is chosen as the tree-based model, the model is constructed as following and attach as 'decisionTree.png'. The maximum depth of the tree is 6 since the tree size is big.

Figure 19 Evaluation Attribute Decision Tree

```
==================
Confusionn matrix
[[19805     5]
 [    2 14433]]
==================
```

Figure 20. ROC curve for Decision Tree



AUC: 0.9997194074570911

For evaluation, accuracy=0.99, recall=0.99, precision=0.93, confusion matrix, and AUC value are calculated. Decision tree algorithm forms a hierarchical structure that each node is set to have some rules with different features so that the features will be more and more specific to using Gini index. Decision tree can be unstable when small variation in the dataset can result in completely different tree, however, it does not need to normalize data and remove missing values, therefore, its performance is really good. Accuracy, recall, precision, and AUC is over 99%. Although the data is preprocessed, the algorithm can ignore some mis-preprocessed procedure. Moreover, running time is 0.3265, which is short enough for decision.

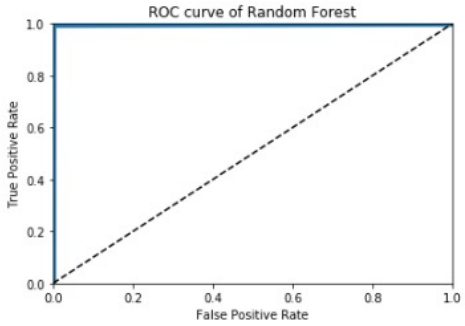
Random Forest Algorithm

Random forest is chosen as ensemble algorithm for the project, it uses multiple decision trees then vote and rank each decision tree so that the best tree with highest accuracy is chosen. Thus, the model by random forest algorithm is similar to the tree model generated by decision tree. The model is constructed as following and attach as 'randomForest.png' for more details. The tree size is big therefore the depth is 6 as following.

Figure 21. Evaluation attribute for RF model

```
==================
Confusionn matrix
[[19804     6]
 [   76 14359]]
==================
```

Figure 22 ROC curve for RF model



AUC: 0.9974174557849274

Accuracy=0.99, recall=0.99, precision=0.93, confusion matrix, and AUC value are calculated. Performance is also great illustrated with accuracy, recall precision and AUC. The weakness of random forest is the running time is 12.75 which is longer than decision tree, but the result of this algorithm is similar to the decision tree algorithm since random forest algorithm needs to generate multiple decision tree to vote for output result.

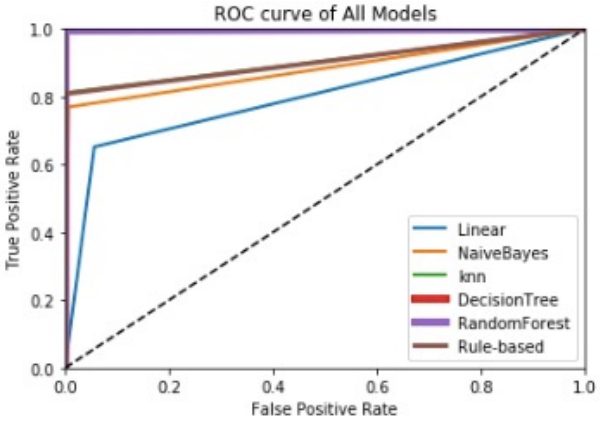
VI. DISCUSSION OF ALGORITHM'S STRENGTH AND LIMITATION

Linear model is used as basic model for observation the evaluation shows that the performance of the classifier is bad. Comparing to other algorithms, the Naïve Bayes performance is fast and running time is lower than 0.1. On the other hand, KNN algorithms shows that the evaluations are higher than 90% but the computing time is long. Also, the time for finding the suitable k value is also long, therefore, the K-nearest Neighbor is not a good method because of the cost is high. Although Naïve Bayes and K-nearest Neighbor are better than linear model, the evaluations present lower accuracy than Decision tree and Random forest. Random forest algorithm is slow in this case since it need to rank and vote multiple decision trees. Figure 23 illustrates a overall comparison of ROC curve of all classifiers used in the project. It gives a visual result for determining the performance of the algorithms. Consider with the result and comparison of ROC curve in Figure 23, decision tree algorithm and random forest algorithm can be chosen, the curve are extremely close to the TPR axis and almost reach the top border of the graph. In addition, the computation cost and running time of decision tree algorithm is less than random forest algorithm, therefore, the decision tree algorithm should be chosen for the result. On the other hand, if running time is an extremely restricted for the consideration, Naïve Bayes can be chosen with fast running time and above average performance. For final decision, decision tree algorithm is chosen after overall comparison among the evaluations of 6 model constructions in table 6.

Table 6. Comparisons of Evaluate attributes among models

	Accuracy	Recall	Precision	Running time
Linear	0.47	0.47	0.71	0.1304
Naïve Bayes	0.89	0.89	0.91	0.0455
K-nearest Neighbor	0.91	0.91	0.92	0.7344
Decision tree	0.99	0.91	0.92	0.3265
Random forest	0.99	0.99	0.99	12.75
Rule based	0.92	0.92	0.92	42.30

Figure 23. ROC curve comparison.



VII. CONCLUSION AND FUTURE WORK

In conclusion, in this project, data preprocess method is implemented with normalization and scaling. Imbalance is achieved by join all class 1 (robot) rows and selected class 0 (human) rows. In total 114148 rows are used from the raw data. Multiple algorithms for the model construction and evaluation such as precision, recall, accuracy, ROC curve, running time are studied and compared in this project to find relationship from the data. And as the result, decision tree algorithm is chosen as the best algorithm for the model construction for good performance for predictions and lower computation cost.

The project can be further developed with improved features, for example, mean value of time, to explore if it can improve the performance. Outliers and noisy data are not eliminated in this project, outlier detection can be used so that more accurate result may generate. In addition, more algorithms can be studied such as bagging classifier, to find better model suitable for this case. Moreover, the project can operate with other projects that study bidding strategy for bidding robots, to improve the detective system for bidding robots occur in online auctions.

Although there are numbers of benefits for users to develop bidding robot to increase winning probability of online auction, the ethical problems cannot be ignored considering the fairness of the trading market. This can be important topic when studying the online bidding.

Reference

[1] D Poole, A. Mackworth, R. Goebel (1998), “Computational Intelligence”, *Oxford University Press*. Available: https://www.cs.ubc.ca/~poole/ci/contents.html Accessed: 18-Nov-2019.

[2] A. Nayak and K. Dutta(2017), "Impacts of machine learning and artificial intelligence on mankind," *2017 International Conference on Intelligent Computing and Control*, Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8321908&isnumber=8321763 Accessed: 19-Dec-2019.

[3]Fei Zhang, Weidong Chen and Yugeng Xi(2005), "Improving Collaboration through Fusion of Bid Information for Market-based Multi-robot Exploration," *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1570272&isnumber=33250 Accessed: 18-Nov-2019.

[4] P. Dasgupta(2012), "Multi-agent coordination techniques for multi-robot task allocation and multi-robot area coverage," *2012 International Conference on Collaboration Technologies and Systems.* Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6261030&isnumber=6261004 Accessed: 18-Nov-2019.

[5] H. Weisbaum(2013), “Bidbots sometimes used to rig Internet penny auction sites,” NBCNews.com. Available: https://www.nbcnews.com/businessmain/bidbots-sometimes-used-rig-internet-penny-auction-sites-1C8673443. Accessed: 18-Nov-2019.

[6] Li Du, Lili Liu and Qian Chen(2011), "Bidding behavior and profits in pay-per-bid auctions," *Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)*, Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6010908&isnumber=6009617 Accessed: 19-Dec-2019.

[7] Zhang Jie and Zhang Yaping(2011), "Research on duration and bid arrivals in eBay online auctions in the internet," *Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)*, Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6010537&isnumber=6009617 Accessed: 18-Nov-2019.

[8] J. Zhou, K. Wang, W. Mao, Y. Wang and P. Huang(2017), "Smart bidding strategy of the demand-side loads based on the reinforcement learning," *2017 IEEE Conference on Energy Internet and Energy System Integration (EI2)* Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8245286&isnumber=8244404 Accessed: 18-Nov-2019.

[9] T. Ito, N. Fukuta, T. Shintani and K. Sycara (2000), "BiddingBot: a multiagent support system for cooperative bidding in multiple auctions," *Proceedings Fourth International Conference on MultiAgent Systems*, Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=858494&isnumber=18605 Accessed: 18-Nov-2019.

[10] M. Berhault *(2003)*., "Robot exploration with combinatorial auctions," *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1248932&isnumber=27959 Accessed: 18-Nov-2019.